CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

The Chromosome-Based Rubber Tree Genome Provides New Insights into Spurge Genome Evolution and Rubber Biosynthesis

Jin Liu[1,2,13], Cong Shi[2,3,13], Cheng-Cheng Shi[4,13], Wei Li[5,13], Qun-Jie Zhang[5,13], Yun Zhang[6], Kui Li[7,8], Hui-Fang Lu[9], Chao Shi[2], Si-Tao Zhu[4], Zai-Yun Xiao[1], Hong Nan[2,3], Yao Yue[4], Xun-Ge Zhu[2,3], Yu Wu[1], Xiao-Ning Hong[4], Guang-Yi Fan[4,9], Yan Tong[2], Dan Zhang[5], Chang-Li Mao[1], Yun-Long Liu[2], Shi-Jie Hao[4], Wei-Qing Liu[9], Mei-Qi Lv[4], Hai-Bin Zhang[2], Yuan Liu[2], Ge-Ran Hu-tang[2,3], Jin-Peng Wang[3,10], Jia-Hao Wang[4], Ying-Huai Sun[9], Shu-Bang Ni[1], Wen-Bin Chen[9], Xing-Cai Zhang[11], Yuan-Nian Jiao[10], Evan E. Eichler[12], Guo-Hua Li[1], Xin Liu[4,9,*] and Li-Zhi Gao[2,5,*]

[1]Yunnan Institute of Tropical Crops, Jinghong 666100, China

[2]Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwestern China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, China

[3]University of Chinese Academy of Sciences, Beijing 100049, China

[4]BGI-Qingdao, Qingdao 266555, China

[5]Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China

[6]Asia-Pacific Tropical Forestry Germplasm Institution, Southwest China Forestry University, Kunming 650224, China

[7]School of Life Sciences, Nanjing University, Nanjing 210023, China

[8]Novogene Bioinformatics Institute, Beijing 100083, China

[9]BGI-Shenzhen, Shenzhen 518083, China

[10]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[11]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

[12]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

[13]These authors contributed equally to this article.

***Correspondence: Xin Liu (liuxin@genomics.cn), Li-Zhi Gao (lgaogenomics@163.com)**

https://doi.org/10.1016/j.molp.2019.10.017

ABSTRACT

The rubber tree, *Hevea brasiliensis*, produces natural rubber that serves as an essential industrial raw material. Here, we present a high-quality reference genome for a rubber tree cultivar *GT1* using single-molecule real-time sequencing (SMRT) and Hi-C technologies to anchor the ~1.47-Gb genome assembly into 18 pseudochromosomes. The chromosome-based genome analysis enabled us to establish a model of spurge chromosome evolution, since the common paleopolyploid event occurred before the split of *Hevea* and *Manihot*. We show recent and rapid bursts of the three *Hevea*-specific LTR-retrotransposon families during the last 10 million years, leading to the massive expansion by ~65.88% (~970 Mbp) of the whole rubber tree genome since the divergence from *Manihot*. We identify large-scale expansion of genes associated with whole rubber biosynthesis processes, such as basal metabolic processes, ethylene biosynthesis, and the activation of polysaccharide and glycoprotein lectin, which are important properties for latex production. A map of genomic variation between the cultivated and wild rubber trees was obtained, which contains ~15.7 million high-quality single-nucleotide polymorphisms. We identified hundreds of candidate domestication genes with drastically lowered genomic diversity in the cultivated but not wild rubber trees despite a relatively short domestication history of rubber tree, some of which are involved in rubber biosynthesis. This genome assembly represents key resources for future rubber tree research and breeding, providing novel targets for improving plant biotic and abiotic tolerance and rubber production.

Key words: rubber tree, rubber biosynthesis, chromosome evolution, whole-genome duplication, domestication

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, CAS.

Liu J., Shi C., Shi C.-C., Li W., Zhang Q.-J., Zhang Y., Li K., Lu H.-F., Shi C., Zhu S.-T., Xiao Z.-Y., Nan H., Yue Y., Zhu X.-G., Wu Y., Hong X.-N., Fan G.-Y., Tong Y., Zhang D., Mao C.-L., Liu Y.-L., Hao S.-J., Liu W.-Q., Lv M.-Q., Zhang H.-B., Liu Y., Hu-tang G.-R., Wang J.-P., Wang J.-H., Sun Y.-H., Ni S.-B., Chen W.-B., Zhang X.-C., Jiao Y.-N., Eichler E.E., Li G.-H., Liu X., and Gao L.-Z. (2020). The Chromosome-Based Rubber Tree Genome Provides New Insights into Spurge Genome Evolution and Rubber Biosynthesis. Mol. Plant. **13**, 336–350.

INTRODUCTION

The rubber tree (*Hevea brasiliensis* (Willd. ex A. Juss.) Muell. Arg.), together with a number of other economically important plant species, such as castor bean (*Ricinus communis* L.), cassava (*Manihot esculenta* Crantz), and Barbados nut (*Jatropha curcas*), belong to the spurge family (Euphorbiaceae). Among ~2500 latex-yielding plants, such as *Eucommia ulmoides* Oliver (Wuyun et al., 2018) and *Taraxacum kok-saghyz* Rodin (Lin et al., 2018), only *H. brasiliensis* produces a commercially viable amount of natural rubber (*cis*-1,4-polyisoprene), making up more than 98% of the world's natural rubber production (Backhaus, 1985; Bowers, 1990). These high-quality isoprenoid polymers possess unique physical and chemical properties that are incomparable with any synthetic alternatives (van Beilen and Poirier, 2007). Natural rubber is thus an indispensable source material for numerous rubber products worldwide. In sharp contrast to the environmental pollution of the industrial synthetics, the rubber tree is able to sustainably yield natural rubber that is still imperative for worldwide high-performance engineering components, such as heavy-duty tires. *H. brasiliensis* remains a long-standing target for genetic manipulation in order to improve and industrially enhance the commercial production of natural rubber.

H. brasiliensis is a cross-pollinated tropical tree that grows to 30–40 m tall and can live up to 100 years in the wild (Priyadarshan and Clement-Demange, 2004). Historical records have documented that the domestication of the rubber tree, which is native to the Amazon basin in South America, began in 1896 and then dispersed to Southeast Asia with the transfer of *H. brasiliensis* seedlings, mainly including Malaysia, Indonesia, Thailand, and China today (Chan, 2000). Over a century of traditional breeding has greatly increased rubber productivity from 650 kg ha^{-1} yielded from wild natural populations of *H. brasiliensis* during the 1920s to 2500 kg ha^{-1} harvested in elite cultivars by the 1990s (Priyadarshan and Goncalves, 2003), which is far below the hypothetical yield of 7000–12000 kg ha^{-1} in the rubber tree (Webster and Baulkwill, 1989). Notwithstanding the origin of the rubber tree from the Amazonian basin, the production of natural rubber in South America constitutes merely 2% of the total production worldwide because of the destructive spread of South American leaf blight disease caused by the ascomycete *Microcyclus ulei* in the 1930s (Lieberei, 2007). Thus, although human selection efforts have brought a slow increase in rubber productivity, it is even more urgently required to raise new *H. brasiliensis* varieties with desirable traits, including tolerance to cold, drought, and wind, and particularly resistance to diseases caused by pathogenic fungi. Nevertheless, the domestication from a small number of wild individuals originating from the Amazon basin of South America created a severe bottleneck that has led to a restricted gene pool for cultivated rubber tree germplasms. This stands in contrast to the wealth of phenotypic diversity and genetic adaptations residing in the natural wild population. These have the potential to be exploited through breeding programs that would help expand genomic diversity, important for the generation of more environmentally resilient and high-yielding varieties. The widespread application of marker-assisted selection promises to advance the breeding efficiency but requires access to a high-quality rubber tree genome sequence in order to accelerate the pace at which the untapped reservoir of agronomically important genes can be exploited from the wild.

Considering the tremendous economic importance of the rubber tree, there has been a long-standing effort to obtain a high-quality reference genome assembly (Rahman et al., 2013; Lau et al., 2016; Tang et al., 2016; Pootakham et al., 2017). The first draft genome sequence of the *RRIM600* clone, generated by a Malaysia research team, was very fragmented but provided our first glimpse of the genome structure (Rahman et al., 2013). The second draft rubber tree genome of the cultivar *Reyan7-33-97* was reported based on sequence data from both whole-genome shotgun sequencing (WGS) and pooled BAC clones. While this assembly was significantly improved with respect to annotation, the use of short Illumina reads posed difficulties in resolving highly heterozygous and repetitive DNA sequences (Tang et al., 2016). The *RRIM 600* genome assembly was subsequently improved based on ~155-fold combined coverage with Illumina and PacBio sequence data. As part of that effort, 100 SMRT (single-molecular long-read sequencing) cells were sequenced using a 10-kbp SMRTbell library yielding 45.25 Gb (~21-fold coverage) with an average read length of 6852 bp (Lau et al., 2016). Until recently, deep-coverage 454/Illumina short-read and Pacific Biosciences (PacBio) long-read sequence data were combined to generate a *de novo* hybrid assembly of the preliminary *BPM24* rubber tree draft genome, which was subsequently scaffolded using a long-range "Chicago" technique to obtain the best assembly of 1.26 Gb (N50 = 96.8 kb) to date. Using a single-nucleotide polymorphism (SNP)-based genetic map, only ~28.9% of the genome assembly (~363 Mb) was successfully anchored into rubber tree's 18 linkage groups (Pootakham et al., 2017). Notwithstanding the complexity of the rubber tree genome, a high-quality chromosome-level genome assembly is required to provide a comprehensive understanding of the latex biosynthesis, genetic improvement of desirable agronomic traits, and efficient utilization of introduced alleles from wild populations to expand the genetic basis of the cultivated rubber tree.

Here, we present a high-quality genome sequence of the rubber tree cultivar *GT1*, an elite cultivar cultivated worldwide, based on

the SMRT technology. This assembly has a length of 1.47 Gb and spans ~94% of the estimated genome size of 1.56 Gb. We anchored this assembly into 18 chromosomes and generated the first chromosome-scale genome assembly assisted by Hi-C technology. Together with the data analysis of resequenced genomes and comparative transcriptomics of cultivated and wild rubber trees, we provide novel insights into gene and genome evolution, domestication, and latex biosynthesis in the rubber tree.

RESULTS

Genome Sequencing, Assembly, and Feature Annotation

We first performed a WGS analysis of a rubber tree cultivar *GT1* with the next-generation sequencing (NGS) Illumina HiSeq 2000 platform. This generated clean sequence datasets of ~348.14 Gb and yielded approximately 261.4-fold coverage (Supplemental Table 1). Using a 17-mer analysis method, we estimated the genome size to be ~1.56 Gb (Supplemental Table 2 and Supplemental Figure 1). The genome heterozygosity was estimated to be 1.60%–1.62% using GenomeScope (Vurture et al., 2017). The genome was assembled using SOAP*denovo* (Li et al., 2010), resulting in the final assembly of ~1.59 Gb (Supplemental Table 3). The contig and scaffold N50 lengths were ~8.79 kb and ~31.3 kb, respectively (Supplemental Table 3).

To resolve the repetitive structure and heterozygous regions (Rahman et al., 2013; Lau et al., 2016; Tang et al., 2016; Pootakham et al., 2017), we also sequenced the same *GT1* individual using the PacBio SMRT sequencing platform. We generated a total of ~161.86 Gb (103.75-fold sequence coverage) of long-read sequence data from 20-kb and 40-kb insert libraries with subread N50 lengths of 9.07 kb and 18.34 kb, respectively (Supplemental Table 4 and Supplemental Figure 2). We subsequently performed a PacBio-only assembly using an overlap layout-consensus method implemented in FALCON (version 0.3.0) (Chin et al., 2013), and obtained a 1.47-Gb genome assembly with a contig N50 of 152.7 kb (Table 1; Supplemental Tables 5 and 6). This final assembly of the rubber tree genome comprises 16 023 scaffolds, of which 15 885 scaffolds (>10 kb) represent 99.93% of the 1.47-Gb genome (Supplemental Tables 5–7). We employed ~61.4 Gb high-quality NGS data with 39.3-fold genome coverage using the Illumina HiSeq 2500 platform to polish the assembled genome (Supplemental Table 4). Next, ~119.63 Gb of sufficiently high-quality Hi-C data (76.69×) were used to further superscaffold the genome assembly (Supplemental Table 8 and Supplemental Figure 3). We finally obtained a reference genome of the rubber tree on the chromosome level by anchoring ~1442-Mbp-sized contigs into 18 chromosomes using AllHic v0.8.12 (Zhang et al., 2019) (Figure 1; Supplemental Tables 9 and 10; Supplemental Figure 4). The total length of the assembled genome sequences accounts for ~92.4% of the estimated genome size (Table 1), about four times longer than the previously reported genome assembly of *BPM24* that was linked using a high-density SNP-based genetic map (Pootakham et al., 2015). The lengths of 18 chromosomes of the *GT1* genome ranged from ~36 Mbp (Chr17) to ~104 Mbp (Chr09) with an average size of ~80 Mbp (Figure 1; Supplemental Table 10; Supplemental Figure 4).

To evaluate the quality of the assembled rubber tree genome, we first mapped ~61 Gbp of Illumina short reads. Our results showed that more than 98% of NGS reads could be unambiguously represented with an expected insert size distribution, indicating a high confidence of genome scaffolding (Supplemental Table 11). We then aligned 51 701 expressed sequence tags retrieved from public databases and 102 235 unigenes assembled through RNA sequencing (RNA-seq) data of the six *GT1* tissues with our assembled genome, showing that ~97.24% and 97.14% of protein-coding genes could be covered, respectively (Supplemental Tables 12 and 13). Finally, we assessed core gene statistics using BUSCO (Simao et al., 2015) to verify the sensitivity of gene prediction and the completeness and appropriate haplotig merging of the genome assembly. Our predicted genes recovered 1359 of the 1440 (94.4%) highly conserved core proteins in the Embryophyta lineage (Supplemental Table 14).

Whole-genome comparisons of the *GT1* genome with the three other rubber tree genome assemblies showed that the ordered syntenic genomic sequences displayed a good collinearity between *GT1* and the three other rubber tree genomes; ~89%–95% of the *GT1* genome sequences with lengths of >1 kbp were covered by any of the other three genomes (Supplemental Table 15; Supplemental Figures 5 and 6). However, the *GT1* genome assembly using long-read PacBio sequencing data alone in this study revealed closer syntenic relationships with hybrid assemblies of *RRIM600* and *BMP24* with Illumina and PacBio sequencing data (Lau et al., 2016; Pootakham et al., 2017) than the fairly fragmented genome assembly of *Reyan7-33-97* using the Illumina sequencing data alone (Tang et al., 2016). We annotated approximately 1042.42 Mbp (~70.82%) of repetitive sequences, among which the *GT1* genome comprised ~65.88% (~969.72 Mbp) of long terminal repeat (LTR) retrotransposons (Supplemental Table 16). Total transposable-element content of the *GT1* genome is larger than that of the formerly reported genome assemblies of the two rubber tree cultivars (*Reyan7-33-97*: 66.46%; *RRIM600*: 69.80%) and comparable with *BPM24* (*BPM24*: 71.10%) (Supplemental Table 16), indicating a high-quality performance to *de novo* assemble genomic regions containing highly repetitive sequences using long PacBio reads (Supplemental Figure 6).

Combining *ab initio* prediction and transcriptome sequence alignments from RNA-seq data of the six tissues (Supplemental Tables 17 and 18), we predicted a total of 44 187 protein-coding genes with an overall support of 96.76% (Table 1; Supplemental Tables 19 and 20; Supplemental Figure 7A). Of them, 79.34%, 75.80%, 94.70%, and 96.37% could be functioned with SwissProt, PFAM, TrEMBL, and Interpro databases, respectively (Supplemental Table 20 and Supplemental Figure 7B). We further performed homology searches and annotated non-coding RNA (ncRNA) genes (Supplemental Table 21), yielding 945 transfer RNA (tRNA) genes, 93 ribosomal RNA (rRNA) genes, 61 small nucleolar RNA (snoRNA) genes, 396 small nuclear RNA (snRNA) genes,

Assembly	
Illumina/PacBio sequencing depth (×)	261.38/103.69
Estimated genome size (Mb)	1561
Chromosome number (2*n* = 2*x* =)	36
Total length of assembly (Mb)	1472
No. of contigs	16 023
Contig N50 (kb)	152.7
No. of scaffolds	600
Contig number on chromosomes	15 324
Chromosome length (Mb)	1442
GC content of the genome (%)	33.87
Annotation	
No. of gene models	44 187
Average gene length (bp)	3918.4
Mean exon length (bp)	222.08
Average exon per gene	5.13
Mean intron length (bp)	672.1
No. of functions annotated	42 758
tRNAs	945
rRNAs	93
snoRNAs	61
snRNAs	396
miRNAs	373
Repeat sequence length (Mb)	1042.4
Percentage of repeat sequence (%)	70.81

Table 1. Statistics of the Genome Assembly and Gene Annotation for Rubber Tree.

and 373 microRNA (miRNA) genes. We annotated ~689 255 simple sequence repeats, which will provide valuable genetic markers to assist future genetic improvement of the rubber tree (Supplemental Table 22 and Supplemental Figure 8).

Chromosomal Evolution after a Common Paleotetraploidy in the Spurge Family

Chromosome-scale genome assembly ensures the detection of whole-genome duplication (WGD) events that affected chromosomal evolution in the rubber tree. We examined the rubber tree genome for homologous genomic segments based on sequence similarities of paralogous genes. A total of 1901 syntenic blocks containing 26 403 paralogous gene pairs were identified in the rubber tree genome (Supplemental Table 23). Comparative analyses of paralogous gene pairs from these intragenomic homologous blocks unquestionably showed that rubber tree has experienced two paleotetraploidization events corresponding to sequence divergence peaks at ~0.124 and 0.530 synonymous transversions per site, respectively (Supplemental Figure 9). Comparisons among paralogs and orthologs of the five Malpighiales species show that a recent paleotetraploidy event occurred prior to the split of the *Hevea* and *Manihot* lineages, and an ancient eurosid WGD was shared by all these examined species (Supplemental Figure 9). This finding strongly supports the hypothesis that a recent paleotetraploidy event took place before the divergence of the *Hevea* and *Manihot* species but after the split of the castor bean (Pootakham et al., 2017).

The rubber tree contains the same number of chromosomes (2*n* = 36) as the cassava, which is almost twice as many as that of castor bean (2*n* = 20) (Chan et al., 2010). The genus *Hevea* is closest to *Manihot* in the spurge family (Euphorbiaceae), and they diverged from each other approximately 36 million years ago (Mya) (Bredeson and Lyons, 2016; Pootakham et al., 2017). The chromosome-based genome assembly we obtained for *GT1* permits us to identify the 1901 syntenic blocks of the rubber tree genome by interchromosomal comparisons, within which 26 403 paralogous gene pairs are spread across the 18 chromosomes, falling into the five 2-by-2 chromosome pairs and two 4-by-4 chromosome groups (Supplemental Table 23; Figures 1 and 2A; Supplemental Figure 10). We similarly detected 38 833 orthologous gene pairs corresponding to 1829 conserved syntenic blocks between the rubber tree and cassava genomes (Supplemental Table 23 and Supplemental Figure 11). These can also be divided into five 2-by-2 pairs and two 4-by-4 groups (Supplemental Table 23 and Supplemental Figure 12). (Bredeson and Lyons, 2016) previously reported a similar pattern of conserved synteny comprising five groups

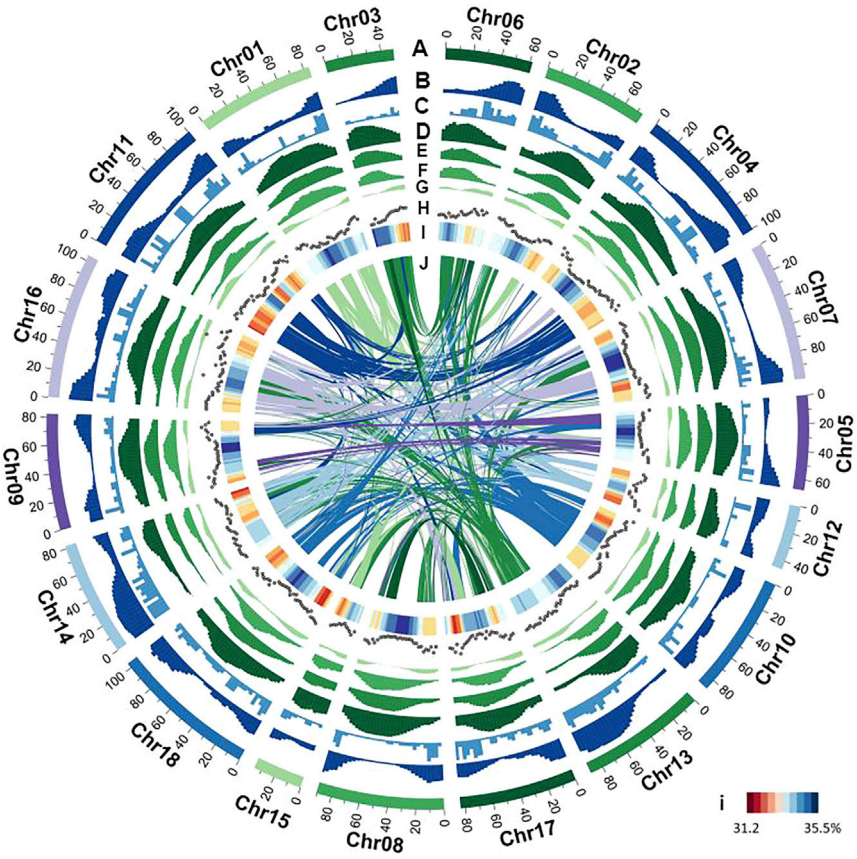

Figure 1. Features of the Rubber Tree Genome.

(A) Circular representation of the 18 pseudo-chromosomes; **(B)** the density of genes; **(C)** the density of non-coding RNA; **(D)** the distribution of transposable elements (TEs); **(E)** the distribution of *gypsy*-type retrotransposons; **(F)** the distribution of *copia*-type retrotransposons; **(G)** the distribution of DNA transposons; **(H)** SSR density; **(I)** the distribution of GC content; **(J)** whole-genome duplication (WGD) event shown by syntenic relationships among duplication blocks containing more than 15 paralogous gene pairs.

of 2-by-2 chromosome pairs and two chromosome groups of 4-by-4 as a result of the paleotetraploidy in the cassava. These two chromosome-based high-quality genome assembles ensured a reliable discovery of macrosynteny conservation between the two euphorbs. Macrosynteny conservation of rubber tree and cassava that comprise the same chromosome numbers further supports the hypothesis that a WGD event occurred in the common ancestor of *Hevea* and *Manihot*.

To trace the palaeohistory of euphorbs and understand the chromosomal evolution after the polyploidization event, we performed a comparative genomic analysis of the rubber tree with cassava (Bredeson et al., 2016) using the grape (Jaillon et al., 2007) as the closest modern representative of the ancestral eudicot karyotype (AEK) (Salse, 2016; Badouin et al., 2017). Since the recent WGD event occurred before the split of the rubber tree and cassava lineages, these two paleopolyploid plants have experienced diploidization through structural and functional changes, providing an unprecedented opportunity to understand chromosomal evolution in spurge plants. We reconstructed a model to infer the scenario of chromosomal evolution in rubber tree and cassava based on genome assemblies at the chromosome level (Figure 2B). In the rubber tree genome, Hbr04, Hbr08, Hbr10, and Hbr13 did not experience many rearrangements while other chromosomes have suffered from a large number of fission and fusion events (Figure 2B). In the cassava genome, Mes03 and Mes04 retained few regions derived from the eudicot ancestor compared with other chromosomes (Figure 2B). Our comparative genomic analysis of the rubber tree and cassava further identified a large number of genomic structural variation after a shared polyploidization event (Supplemental Figure 13). The model of chromosomal evolution for rubber tree and cassava reveals that substantial genomic rearrangement events have extensively shaped the chromosome structure, leading to the 18 modern chromosomes since the common paleopolyploid ancestor.

The Three LTR-Retrotransposon Families Are Drivers of the Expanded Rubber Tree Genome

Using our chromosome-based genome assembly, we investigated the evolution of LTR retrotransposons and their potential contribution to the growth of the rubber tree genome. The rubber tree has experienced a rapid growth of genome size as a result of the *Hevea*-specific proliferation of transposable elements compared with the three other closely related species of Euphorbiaceae, namely cassava, castor bean, and physic nut (Table 1; Figure 3; Supplemental Tables 24 and 25; Supplemental Figures 14 and 15). Retrotransposons in the rubber tree genome are particularly abundant (~991.73 Mb; ~67.38%) compared with five other plant species of Malpighiales, namely *Manihot esculenta*, *Jatropha curcas*, *Ricinus communis*, *Populus trichocapa*, and *Linum usitatissimum*. Specifically, the rubber tree genome is enriched in LTR retrotransposons (Ty1/*copia*, Ty3/*gypsy*, and non-autonomous LTR retrotransposons) with a total length of ~969.72 Mb (~65.88%) (Figure 3; Supplemental Figures 15 and 16; Supplemental Tables 24 and 25). Dating transposable elements shows that retrotransposons and DNA transposons were almost separately amplified among the six sequenced plant species of Malpighiales (Supplemental Figure 14). Comparative analyses of the Ty1/*copia*, Ty3/*gypsy*, and other classes of LTR retrotransposons suggest that they have experienced three retrotransposition bursts over the last 10, 20, and 30 Mya, respectively (Supplemental Figure 14). Ty3/*gypsy* LTR-retrotransposon families dominate the rubber tree genome, contributing ~585.69 Mb (~39.79%), and are ~3.13-fold more abundant than Ty1/*copia* with ~187.34 Mb (~12.73%) (Supplemental Tables 24 and 25; Figure 3A; Supplemental Figure 15). We exclusively observed that the amplification of *Tekay* (~430.60 Mb; ~29.25%) of Ty3/*gypsy*

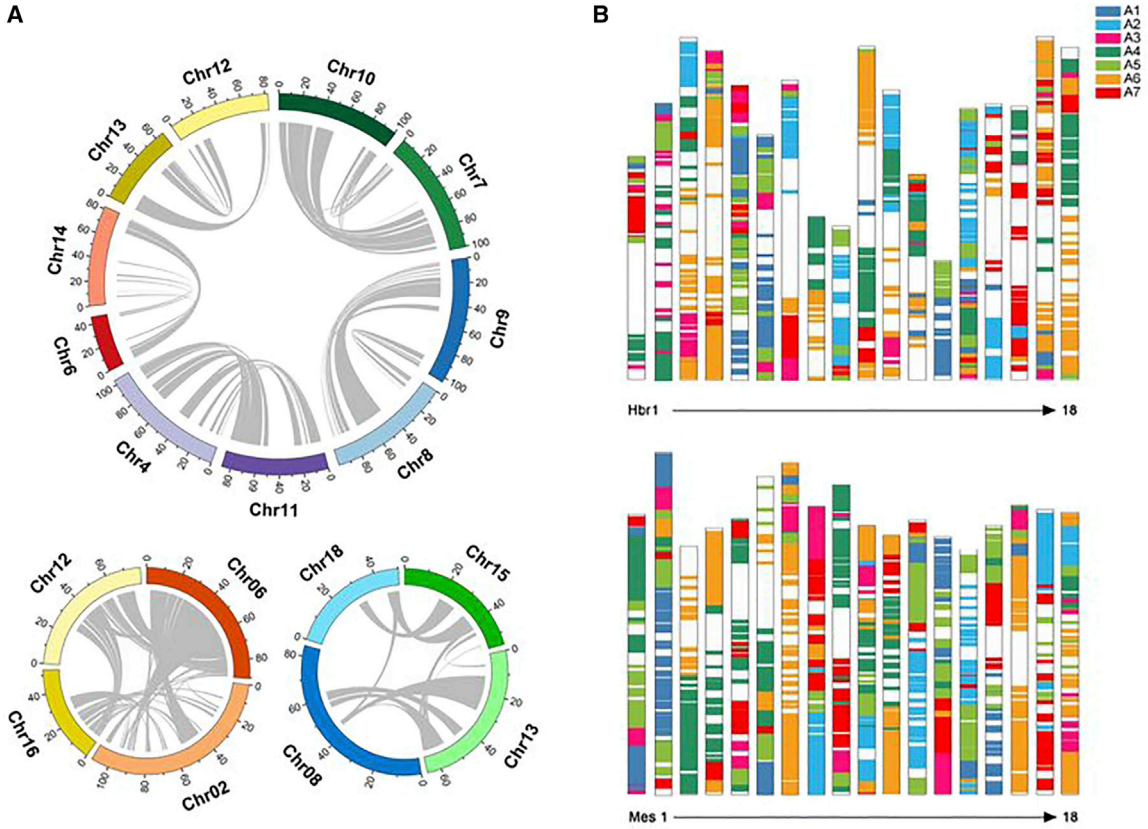

Figure 2. Chromosome Evolution after the Shared Paleotetraploidy of Rubber Tree and Cassava.

(A) Conserved synteny within the rubber tree genome. Diagrams show genomic collinearity within the *H. brasiliensis* genome. Lines link the position of paralogous gene sets among linkage group/chromosome set. The 10 chromosomes arranged in the upper circle illustrate 1:1 synteny between the five duplicated pairs of chromosomes. The eight chromosomes depicted in the lower circle each share syntenic regions with two other chromosomes, owing to chromosomal rearrangements that occurred after the whole-genome duplication.

(B) Evolutionary patterns for the rubber tree and cassava chromosomes from the ancestral eudicot karyotype (AEK) of seven (pre-whole-genome triplication of eudicots) protochromosomes. Genome rearrangements of these two genomes are elucidated with different colors that represent the origins from the seven ancestral chromosomes from the $n = 7$ AEK. Rubber tree linkage groups are designated by "Hbr" followed by the linkage group numbers, and cassava chromosomes are designated by "Mes" followed by the chromosome numbers.

and *Angela* (~96.70 Mb; ~6.57%) of Ty1/*copia* has largely contributed to the expansion of the rubber tree genome (Figure 3; Supplemental Figure 15A and 15B). Surprisingly, the largest Ty3/*gypsy* retrotransposon family *HBL001* (~368.08 Mb; ~25.01%) belonging to *Tekay* has long been active over the last 45 million years, (myr). whereas, the two retrotransposon families, *HBL002* (~96.70 Mb; ~6.57%) and *HBL003* (~62.52 Mb; ~4.25%) belonging to *Angela* and *Tekay*, respectively, have recently proliferated over the last 10 MYR (Supplemental Figures 15C and 17). It is these three LTR retrotransposon families that have predominantly amplified; they together account for ~54.38% of LTR retrotransposons and ~35.83% of the whole rubber tree genome (Figure 3; Supplemental Figures 15C and 17; Supplemental Table 26). Further annotation and comparative analyses of the repeated elements among the four Euphorbiaceae species (Supplemental Figures 15C and 17), including *M. esculenta*, *J. curcas*, *R. communis*and *H. brasiliensis*, suggests that only the *H. brasiliensis* genome has undergone specific bursts of these three LTR-retrotransposon families during the past 5 MYR. This resulted in the accumulation of a large number of retroelements driving the growth of ~970 Mb of the rubber tree genome after the divergence from the *M. esculenta* lineage around 36 Myr (Supplemental Figure 17).

The Rapid Evolution of Gene Families and Rubber Biosynthesis

Defining the rapidly evolving gene families among flowering plants has been helpful to identify genomic basis underlying physiological changes of metabolite constituents during evolution. We compared the predicted proteomes of the rubber tree, cassava, castor bean, physic nut, poplar, flax, *Arabidopsis thaliana*, and rice, yielding a total of 24,562 orthologous gene families that comprised 225,207 genes (Supplemental Table 27 and; Supplemental Figure 18A). This revealed a core set of 121,256 genes belonging to 7,498 gene clusters that were shared among all eight plant species, representing ancestral gene families (Supplemental Figure 18A). We obtained a total of 1,352 gene clusters containing 4,791 genes unique to the rubber tree, potentially related to biosynthetic processes associated with the production of latex. PFAM functional analysis showed that gene functions are enriched in aspartic acid proteinase gene family (PF16845, $P < 0.05$), which

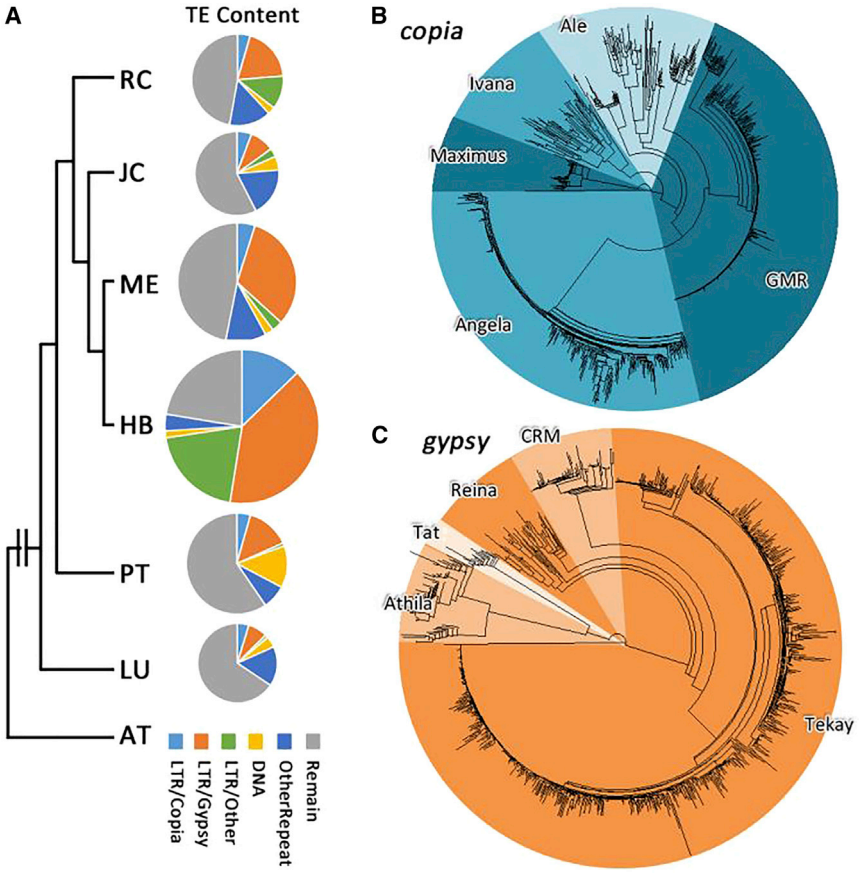

Figure 3. Genome-Size Variation and Evolution of Retrotransposon Families in the Rubber Tree Genome.

(A) shows Genome sizes and proportions of different types of transposable elements (TEs) in *R. communis* (RC), *J. curcas* (JC), *M. esculenta* (ME), *H. brasiliensis* (HB), *P. trichocarpa* (PT) and *L. usitatissimum* (LU) using *Arabidopsis thaliana* (AT) as outgroup.

(B and C) The unrooted phylogenetic trees were constructed on the basis of Ty1/*copia* **(B)** and Ty3/*gypsy* **(C)** aligned sequences corresponding to the RTdomains without premature termination codon. LTR retrotransposon family names and proportion of each are indicated.

encodes enzymes associated with the production of lutoid membranes necessary for the aggregation of rubber particles (Supplemental Table 28). Remarkably, we found that the rubber tree-specific gene families are significantly enriched in functions related to the phosphorylation activity, which is key to biosynthetic processes of latex (PF08645, $P < 0.05$; PF03372, $P < 0.001$) (Supplemental Table 28).

In flowering plants, the expansion or contraction of gene families is an important driver of phenotypic diversification and the formation of phytochemical properties (Ohno, 1970; Chen et al., 2013). We characterized gene families that underwent discernible changes and divergently evolved along different branches with a particular emphasis on those involved in the latex biosynthesis of the rubber tree. Our results revealed that, of the 14,253 gene families inferred to be present in the most recent common ancestor of the eight studied plant species, 5,034 gene families comprising 14,103 genes show significant expansions ($P < 0.05$) in the rubber tree lineage (Supplemental Figures 18B and 19). Functional enrichment analyses of these genes by both gene ontology (GO) terms and PFAM domains surprisingly reveals that they were mainly enriched in a number of functional categories involved in whole latex biosynthesis process that comprises twelve pathways (Rahman et al., 2013) (Supplemental Table 29). For examples, functional annotation of these genes demonstrates that they were mainly enriched in a number of functional categories involved in basal metabolic processes, such as fructose 6-phosphate metabolic process (GO:0006002; $P < 0.001$), glucose metabolic process (PF11721; $P < 0.001$), phospholipase (PF00388, PF12357, PF00614, PF09279, PF00387; $P < 0.001$), ribosomal protein (PF01020, PF00453, PF00347; $P < 0.001$), argonaute (PF08699, PF16487, PF16488, PF16486; $P < 0.001$), aminotransferase (PF00202; $P < 0.001$) and dehydrogenase (PF00725, PF09265; $P < 0.001$). Furthermore, gene families are significantly enriched in a number of functions related to carbohydrate metabolism, such as hydrolase (PF00702, PF03662, PF07748, PF01915, PF01738, PF03644, PF00332, PF01074, PF12215, PF04685, PF00933, PF00232, PF01301, PF12710; $P < 0.001$), glutamine synthetase (PF03951, PF00120; $P < 0.001$), carbohydrate-binding protein of the ER (PF12819; $P < 0.001$), glutamate decarboxylase (GO:0004970; $P < 0.001$), and UDP-glucose/GDP-mannose dehydrogenase family (PF00984, PF03720; $P < 0.001$) (Supplemental Table 29). In addition, gene families are significantly enriched in a number of functions related to pyruvate metabolism involved in the MVA pathway (PF02887; $P < 0.001$). Notably, we find that gene families encoding *S*-adenosyl-L-methionine synthase (SAMS) are significantly enriched in the ethylene biosynthesis, including *S*-adenosylmethionine biosynthetic process (GO:0006556; $P < 0.001$), methionine adenosyltransferase activity (GO:0004478; $P < 0.001$) *S*-adenosylmethionine synthetase (PF02773, PF00438, PF02772; $P < 0.001$), and adenosylmethionine decarboxylase (PF01536; $P < 0.001$). Gene families are also significantly enriched in a number of functions related to the biosynthesis of metabolic compounds, including polysaccharide (PF03033, PF05686, PF00982, PF13641; $P < 0.001$) and glycoprotein lectin (PF00139, PF02140, PF01453; $P < 0.001$) (Supplemental Table 29). Enrichment analyses show that the expanded gene families are enriched in a total of 18 KEGG pathways, including fifteen metabolism-related pathways, such as pyruvate metabolism (ko00620; $P < 0.05$) relevant to biosynthetic processes of latex (Supplemental Table 30 and; Supplemental Figure 18C).

As an important biological feature of the rubber tree, rubber is sequentially synthesized by twelve pathways, in which hundreds

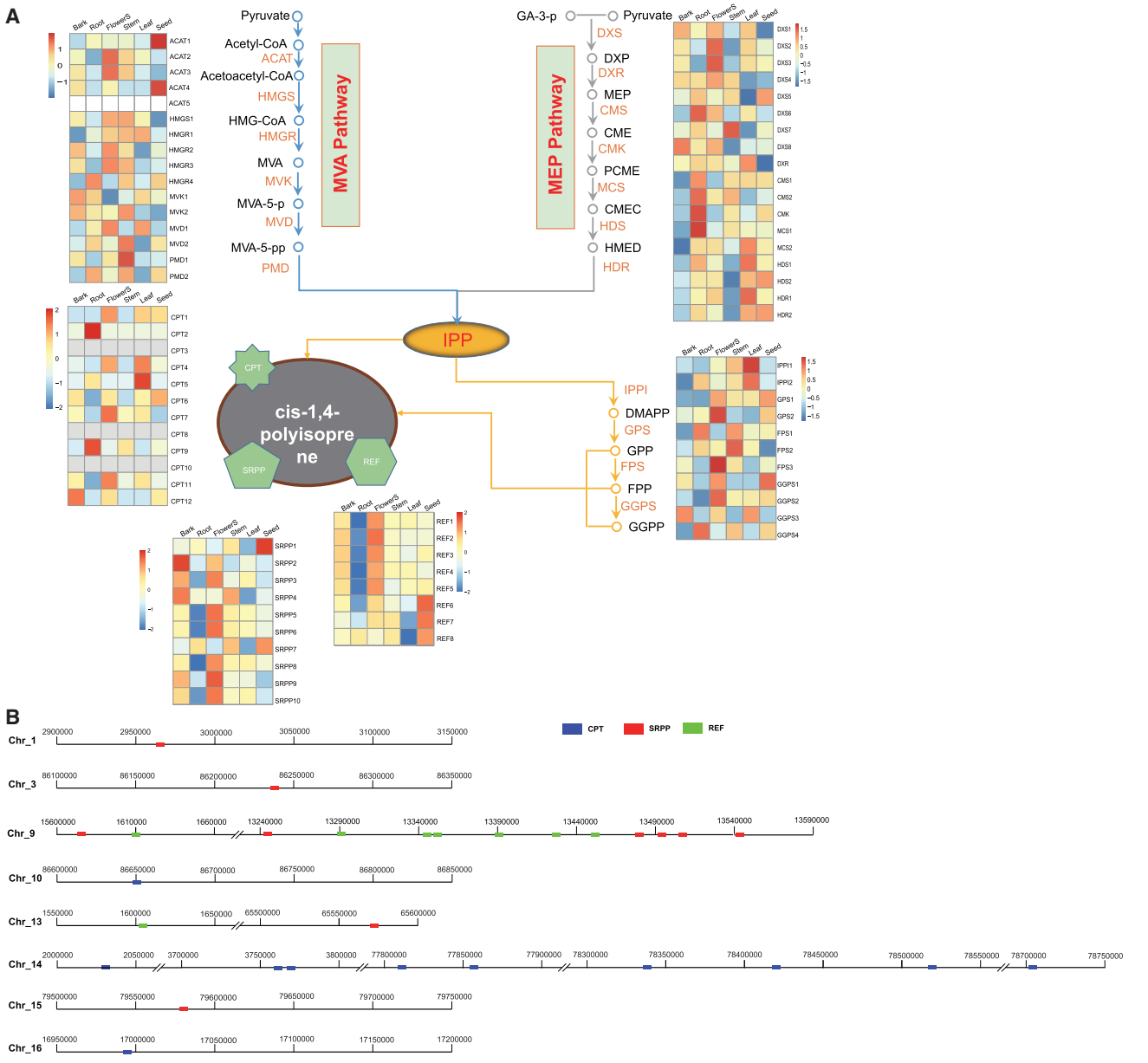

Figure 4. Rubber Biosynthesis and the Expansion of the *REF*/*SRPP* and *CPT* Gene Families in Rubber Tree.

(A) Levels of expression (reads per kilobase per million reads mapped; RPKM) of genes involved in the rubber biosynthesis in bark, root, flower, stem, leaf and seed.

(B) Genomic locations of *REF*/*SRPP* and *CPT* genes. Chromosomes are represented as solid bars with their names on left. Note that most of the *REF*/*SRPP* and *CPT* genes are located on Hbr_9 and Hbr_14, respectively.

of gene families are involved (Rahman et al., 2013). In *H. brasiliensis*, natural rubber is a high molecular weight biopolymer that comprises *cis*-isoprene units resulting from isopentenyl diphosphate (IPP). The biosynthesis of IPP takes place via two distinct paths, that is, MVA and MEP pathways. In total, we identified 70 so-called rubber biosynthesis-related genes, including 11 genes involved in the MVA pathway, 18 genes associated with MEP pathway, 11 genes responsible for initiator synthesis in the cytosol, and 30 genes related to the rubber elongation (Supplemental Table 32 and; Figure 4). Compared to non-rubber-produced plant species, the rubber tree shows the largest number of genes involved in the synthesis of IPP to the final rubber polymer, which may largely enhance the formation of isoprenoids (Supplemental Table 33). We particularly observed a significant expansion of rubber biosynthesis-related gene families that correlates with its capacity to produce high levels of latex in *H. brasiliensis*, such as the eight *f1-deoxy-D-xylulose 5-phosphate synthase* (*DXS*) encoding DXP synthase that catalyzes the first step in the MEP pathway, twelve *cis-prenyltransferase* (*CPT*), eight *rubber elongation factor* (*REF*) and ten *small rubber particle* (*SRPP*) genes (Supplemental Table 33). Besides a significant expansion (13/18) of the *REF*/*SRPP* gene

family that make gene clusters on chromosome 9 (Figure 4), we also found that of the identified twelve *CPT* genes are clustered on chromosome 14 (Figure 4), suggesting that many of them appear to have arisen by tandem duplication events.

To gain insights into the molecular mechanisms underlying the production of natural rubber we examined tissue-specific expression patterns of genes involved in the rubber biosynthesis using RNA-seq datasets from barks, roots, flowers, stems, leaves, and seeds. Our results show that these 20 rubber biosynthesis-related gene families exhibit distinct patterns of gene expression among tissues; interestingly, they are most highly expressed in flowers when compared to barks, followed by stems (Supplemental Table 32). We paid particular attention to tissue-specific expression patterns of *REF/SRPP* and *CPT* gene families that are functionally known to encode enzymes that are responsible for the rubber elongation. The 18 *REF/SRPP* genes exhibit dissimilar expression patterns among six tissues of the rubber tree, of which we intriguingly found that *REF1*, *REF2*, *REF3*, *REF4*and *REF5* are highly expressed in flowers, while *REF6*, *REF7*and *REF8* are highly expressed in seeds. However, the majority of *REF* genes are expressed in barks. Interestingly, we observed that *SRPP3*, *SRPP5*, *SRPP6*, *SRPP8*, *SRPP9*, and *SRPP10* are highly expressed in flowers compared to the barks, while *SRPP2* and *SRPP4* are favorably expressed in barks when compared to flowers. Our results indicated that *CPT1*, *CPT4*, *CPT7*, and *CPT11* are highly expressed in flowers, while others are preferentially expressed in barks (*CPT12*), roots (*CPT2* and *CPT9*), leaves (*CPT4* and *CPT5*) and seeds (*CPT6*), respectively. Taken together, we show that, besides the high expression of a small number of genes in roots, stems, leaves, and seeds, the rubber may be actively synthesized in both flowers and barks, but predominantly accumulates in barks. Our differential expression analysis further reveals that 276 DEGs related to the rubber biosynthesis are commonly up- regulated in barks compared with five other tissues (Supplemental Table 34 and; Supplemental Figure 20), of which we identified two latex-produced genes involved in TCA-cycle and one gene associated with starch metabolism (Supplemental Table 35).

To provide a transcriptomic overview of expression divergence present between cultivated and wild rubber trees that underlies the variation in latex yields, we examined tissue-specific expression patterns of genes involved in the rubber biosynthesis using RNA-seq data from barks of the five elite cultivars of the rubber tree and the five accessions of wild rubber tree representatives of its global geographic range. Our results show that these 20 rubber biosynthesis-related gene families exhibited distinct patterns of gene expression across different accessions of rubber trees (Supplemental Figure 21 and; Supplemental Tables 36, 37, and 38). PCA analysis based on expression levels of rubber biosynthesis-related genes further suggests that, compared to abundant expression bias across *GT1* tissues, there are no remarkably preferential expression patterns between the cultivated and wild rubber trees (Supplemental Figures 21 and 22; Supplemental Tables 36 and 37). Our statistic test further showed that, among all 69 rubber biosynthesis-related genes, only seven genes are significantly differentially expressed between cultivated and wild rubber trees (Supplemental Table 38), which were experimentally validated by real-time PCRs (Supplemental Figure 23 and; Supplemental Table 39).

Genomic Insights into Population Divergence and Artificial Selection on Cultivated Rubber Tree

The first high-quality rubber tree genome allows us to examine genomic variation present within the cultivated and wild rubber trees and detect potential signatures of selection during the domestication. We employed the Illumina short-read technology with paired-end libraries on the HiSeq2000 sequencing platform to resequence eight elite cultivars of the rubber tree and six accessions of wild rubber tree according to native geographic range throughout the world (Supplemental Table 40). Each of these 14 cultivated and wild representative genomes (first reported in this study), were sequenced to >3-fold sequence coverage using paired-end (2 × 100-bp) Illumina sequencing. In addition, we included two previously reported genomes of cultivated rubber trees (RRIM 600 and Reyan7-33-97) with >10-fold genome sequence coverage (Lau et al., 2016; Tang et al., 2016). We first mapped high-quality short reads back to the reference genome sequence of the rubber tree, and obtained mapping rates ranging from 95.81% to 98.86% (Supplemental Table 40). They represent the largest whole-genome sequencing data for a diverse panel including 16 accessions with sufficient depths of genome coverage to represent the genomic diversity of cultivated and wild rubber trees.

After aligning reads against the rubber tree reference genome sequence, we obtained a total of 15,728,276 common single nucleotide polymorphisms (SNPs), more than 90% of which were located on intergenic regions (Supplemental Tables 40 and 41; Supplemental Figure 24). The density of SNPs across all 16 cultivated and wild rubber trees averages approximately 11.9 SNPs per kilobase, of which 14,645,744 and 14,497,454 SNPs were totally identified across these 10 cultivars and 6 wild accessions of the rubber tree, respectively. We observed 1,547,989 SNPs (9.84%) in genic regions, including 174,394 synonymous, 258,938 nonsynonymous, 426,705 exonic, and 1,121,284 intronic SNPs. We used this large dataset of SNPs from both wild and cultivated rubber trees to evaluate genome-wide levels of nucleotide diversity (π and θw) to be 0.00345 ± 0.00021 and 0.00291 ± 0.00041. The wild rubber trees show higher levels of nucleotide diversity than rubber tree cultivars (π per kb, 3.43 *vs*. 2.86, $P = 2.29 \times 10^{-5}$; θw per kb, 2.59 *vs*. 2.55, $P = 4.21 \times 10^{-1}$) (Supplemental Table 42).

The genome-wide SNP dataset provides an unprecedented opportunity to investigate the population structure of the rubber tree. Principal-component analysis (PCA) of SNP variation suggests that the top two eigenvectors strongly correlate with sampling sources: PC1 was correlated with cultivated rubber trees admixed by two accessions of wild rubber tree (Wild258 and Wild166), and PC2 was correlated with wild rubber trees (Figure 5A). These inferred relationships are also supported by phylogenetic analysis based on SNPs (Figure 5B), showing that the ten cultivars (Cult053, Cult169, Cult171, Cult223, Cult293, Cult589, KEN, RRIM 600, RP, and RY7-33-97) formed one cluster admixed by two accessions of wild rubber tree (Wild258 and Wild166), while the four wild rubber trees (Wild162, Wild194, Wild232 and Wild778) were clustered into

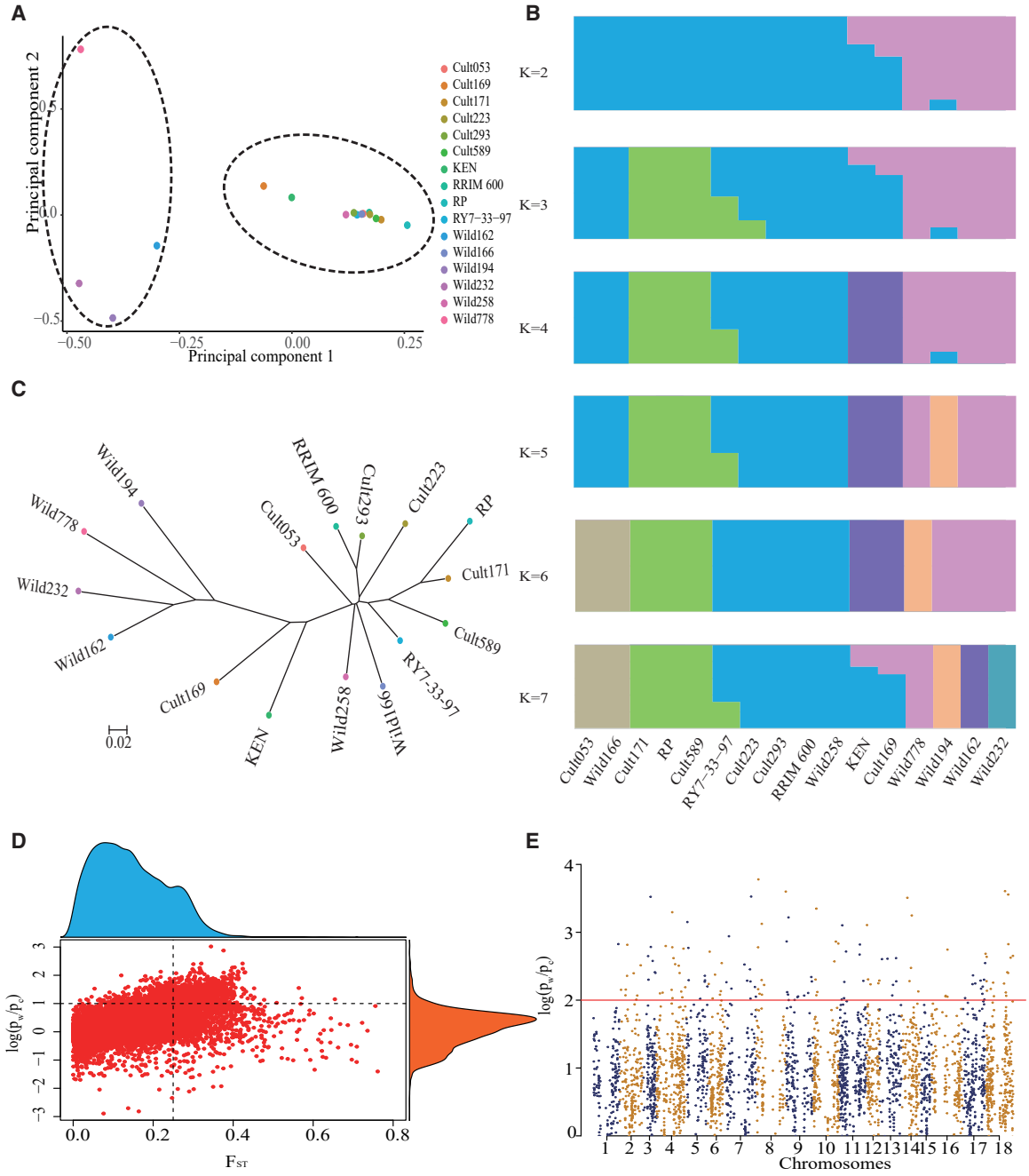

Figure 5. Population Divergence and Artificial Selection of *H. brasiliensis*.

(A) Two-way principal components analysis (PCA) of the sixteen *H. brasiliensis* accessions using identified SNPs.

(B) Population structure of *H. brasiliensis*. Each color and vertical bar represents one population and one accession, respectively. The *y*-axis shows the proportion of each accession contributed from ancestral populations.

(C) Neighbor-joining (NJ) phylogenetic tree of the sixteen *H. brasiliensis* accessions constructed using SNP data. The scale bar represents the evolutionary distances measured by p-distance.

(D) the Distribution of the F_{ST} values and levels of nucleotide variation between the cultutivated and wild rubber trees.

(E) Genomic regions under artificial selection on the 18 chromosomes of rubber tree.

another group. To generate an alternative view of population stratification, we used the population clustering program STRUCTURE (Pritchard et al., 2003), which inferred the optimal number of genetic clusters comprising the cultivated and wild rubber tree genomes to be $K = 7$ (Figure 5C; for other K values, see Supplemental Figure 25). We found that cultivated rubber trees predominantly form three major clusters, two of which are grouped with Wild258 and Wild166, respectively. However, the four accessions of wild rubber trees split into four distinct groups, of which Wild778 has contribute the most to cultivated rubber trees, probably serving as an ancestral population.

Population genomics provides an opportunity to track the dispersal, genetic bottlenecks and signature of artificial selection during the domestication in most cultivated crop species (Doebley et al., 2006). To detect the footprint of artificial selection in the rubber tree genome we attempted to identify genomic regions with significantly lowered levels of polymorphisms in cultivated populations compared to wild rubber trees. While comparing to the F_{ST} values and levels of polymorphisms between cultivated and wild accessions of the rubber tree (Figure 5D and 5E), we performed a whole-genome screening and identified ~83.7 Mbp of the genome potentially subjected to selective sweeps. These regions harbored 578 genes, which we consider candidate domestication genes of the rubber tree (Supplemental Table 43). KEGG enrichment analysis shows that these candidate domestication genes are enriched in twelve pathways, of which nine genes enriched in ko00900 associate with the terpenoid backbone biosynthesis that is key to the latex biosynthesis (Supplemental Figures 26 and 27). Further functional annotation indicates, for example, that the two genes (*GT005609*, *GT009352*) encode the first enzyme (DXS) in the MEP pathway (Supplemental Table 43).

DISCUSSION

De novo sequencing and assembly of large, highly repetitive, and heterozygous plant genomes have long been challenging and problematic. Here we construct a high-quality reference for one such genome, the *GT1* genome for *H. brasiliensis*. We generated a 1.47-Gbp *de novo* assembly of the rubber tree genome highlighting the potential of combining PacBio long-read and Illumina short-read assemblies to create improved reference genomes, in sharp contrast to relatively fragmented genome assemblies using the Illumina sequencing data alone (Rahman et al., 2013; Tang et al., 2016) or hybrid assemblies with Illumina and PacBio sequence data (Lau et al., 2016; Pootakham et al., 2017). We thus established the efficiency of employing long SMRT reads to resolve ambiguous genomic regions harboring predominantly repetitive sequences, particularly LTR-retrotransposons.

Considering the difficulty in generating a high-density linkage map for the rubber tree as a long lifespan tree species, we demonstrated an efficient methodology that leverages long-range Hi-C data to scaffold contigs assembled from PacBio reads and successfully anchor ~98% of the 1.47-Gb genome assembly into 18 pseudo-chromosomes. After the availability of four draft genome assemblies for the rubber tree (Rahman et al., 2013; Lau et al., 2016; Tang et al., 2016; Pootakham et al., 2017), we obtain a chromosome-based reference genome with considerable improvement in sequence contiguity. The advent of the *GT1* reference genome together with companion transcriptome resources presented in this study provides important insights into spurge genome evolution. It also further strengthens interest in the rubber tree as a model for ~2,500 rubber-produced plants to accelerate our understanding of the molecular mechanisms underlying the rubber biosynthesis.

Our comparative genomics analysis of the five Malpighiales species convincingly demonstrates that the rubber tree has experienced two paleotetraploidization events. Besides an ancient eurosid WGD (Salse, 2016), we confirm a recent paleotetraploidy event occurred before the divergence of the *Hevea* and *Manihot* species but after the split of the castor bean by genome-scale comparative analysis of *H. brasiliensis* and other members in Euphorbiaceae, *M. esculenta*, *R. communis*and *J. curcas* (Chan et al., 2010; Sato et al., 2011; Bredeson and Lyons, 2016; Pootakham et al., 2017). Chromosome-based analysis of the two high-quality genomes of *M. esculenta* (Bredeson and Lyons, 2016) and *H. brasiliensis* reported in this study further reveal that macrosynteny conservation derived from the shared polyploidization event were disrupted by widespread genomic rearrangement events that drive chromosomal evolution in the spurge family.

The rubber tree possesses an unusually large genome when compared to the majority of other spurge plants. We compared the *GT1* genome with three other previously released genome assemblies, including RRIM600 (Rahman et al., 2013; Lau et al., 2016), Reyan7-33-97 (Tang et al., 2016) and BMP24 (Pootakham et al., 2017), showing nearly completely annotation of most transposable elements, including a large number of LTR retrotransposons in the *GT1* genome. We identify a large *Hevea*-specific proliferation of LTR retrotransposons, which contributed to the rapid increase in genome size when compared to the three other closely related species, including cassava, castor bean and physic nut. Among the three LTR retrotransposon families that have predominantly amplified and altogether account for over half of whole rubber tree genome, we observed a pattern of longstanding and incessant LTR retrotransposon bursts of the largest Ty3/*gypsy* retrotransposon family over the last 45 million years. This is similar to what has been reported for tea tree genome (Xia et al., 2017). However, we also document the recent proliferation of two other large LTR retrotransposon families during the last 10 MYR. Once again, this pattern is consistent to a previous study that reported a recent proliferation of the three LTR-retrotransposon families in the wild rice genome, *Oryza australiensis*, leading to a two-fold increase in genome size during the last three MYR (Piegu et al., 2006).

The well-established WGD event occurred before the divergence of the *Hevea* and *Manihot* species but after the split of the castor bean. This occurred in conjunction with lineage-specific segmental duplication leading to the expansion of gene families relevant to the activation of the latex biosynthesis and thus the increase of rubber production. We have identified a large number of rubber tree-expanded genes encoding enzymes widely involved in numerous functional categories related to basal metabolic processes, carbohydrate metabolism as well as pyruvate metabolism relevant to the MVA pathway, which have greatly enhanced the accumulation of sufficient precursors for the subsequent latex production. The rubber tree-expanded genes encoding SAMS are significantly enriched in GO terms and PFAM domains involved in the ethylene biosynthesis that have greatly improved the latex production (Yang and Hoffman, 2003; Dusotoit-Coucaud et al., 2010). We detected a significant enrichment in a number of functions related to glycoprotein lectin and polysaccharide activities that are well-known to regulate latex coagulation and the blocking of latex flow—important for controlling the rubber production and yield of the rubber tree. The completion of such a high-quality reference genome of rubber tree also permits us to fully identified almost all gene families involved in sequentially long pathways of the

rubber biosynthesis from sucrose to rubber polymerization. Our comparative analyses with non-rubber-produced plant species show that the rubber tree harbors the largest number of genes involved in the latex biosynthesis. The rapid expansion of rubber biosynthesis-related gene families, such as *DXS*, *CPT*and *REF/SRPP*, which is in a good agreement with former observations (Lau et al., 2016; Tang et al., 2016; Pootakham et al., 2017), suggests a strong correlation with its capacity to produce high levels of latex in *H. brasiliensis*. We thus hypothesize that, instead of the assumption that the expansion of the *REF/SRPP* family is associated with correlation with the rubber biosynthesis (Tang et al., 2016), the formation of rubber biosynthesis network and wide-ranging expansion of rubber biosynthesis-related gene families enabled the rubber tree to significantly strengthen the latex biosynthesis and to efficiently yield high-quality natural rubber.

Comprehensive comparative transcriptomic analyses aided by this high-quality genome assembly also provided new insights into the molecular mechanisms underlying the production of natural rubber. An assessment of transcriptomic data shows that these rubber biosynthesis-related gene families are more highly expressed in flowers than barks and stems, suggesting that a potential subfunctionalization or neofunctionalization after gene duplication might explain functional divergence under strong selection pressures that exaggerates the complexity in rubber biosynthesis. We previously reported that the majority of genes involved in the ginsenoside biosynthesis are highly expressed in flowers compared with leaves and roots in *Panax notoginseng* (Zhang et al., 2017). The results suggested that ginsenosides might be actively synthesized in flowers besides roots and leaves but accumulated in roots, supported by a speedy accumulation of total triterpene saponins after flowering. We thus assume that the rubber may be actively synthesized in flowers besides barks but accumulated in barks, but the hypothesis that there is an increased yield of rubber in bark after flowering requires to be experimentally validated in *H. brasiliensis*.

In addition to confirming earlier observations that *REF/SRPP* gene families are highly expressed in latexs (Lau et al., 2016; Tang et al., 2016; Pootakham et al., 2017), we observed differentially expressed patterns of *REF/SRPP* and *CPT* genes in flowers and barks. Further experimental studies are required to examine the complicated transcriptional regulation of these expanded rubber biosynthesis-related gene families. Such investigations will likely offer clues to understanding the unique properties of *H. brasiliensis* needed to produce extraordinary amounts of rubber.

We also provided new insights into natural standing genetic variation and divergence between the cultivated and wild rubber trees. Compared to wild rubber trees, we observe considerable reduction in genomic diversity among cultivated rubber trees. This is likely the result of a strong genetic bottleneck and artificial selection during the relatively short period of domestication since the late nineteenth century. Despite limited phenotypic divergence and gene flow, there is a strong genetic split between wild and the cultivated rubber trees. A subset of wild rubber trees, however, may either represent the ancestral populations of the domesticated rubber tree or have experienced more recent and possibly frequent genomic introgression with rubber tree cultivars. Extensively sampling of cultivated and wild rubber trees will be required to distinguish among these possibilities. Finally, we identify hundreds of candidate domestication genes many of which are involved in specific pathways important for rubber biosynthesis. More analyses and experimental validation are needed to determine whether these represent bona fide domestication genes or whether they are simply genetic hitchhikers of regions that were targeted by artificial selection over one hundred years' domestication of the rubber tree.

The rubber tree genome assembly and transcriptomic and genomic variation datasets presented in this study will offer valuable information to aid efficient germplasm exploration and the improvement of economically important traits of rubber tree to meet global market's increasing demand for natural rubber.

METHODS

Genome Sequencing and Assembly

DNA was extracted from a *GT1* individual for PacBio RSII and HiSeq sequencing platforms. One library with 20 kb insert size was constructed and sequenced for 100 SMRT cells on PacBio RSII. The PacBio data was assembled by FALCON (version 0.3.0) (Chin et al., 2013). We generated ~61 Gb Illumina data with 500 bp insert size on HiSeq 2500 platform to polish assembled genome sequences using pilon (Walker et al., 2014). For Hi-C sequencing, chromosome structure was fixed by formaldehyde crosslinking, and then *MboI* enzyme was used to shear DNA. Hi-C library with 200-600 bp insert size was constructed, which was sequenced on HiSeq 2000 platform. The Hi-C sequence data were qualified with HIC-pro (Servant et al., 2015), in which the validly mapped reads were selected to cluster and order the assembled genome sequences using AllHic v0.8.12 (Zhang et al., 2019).

Genome Annotation

Repetitive elements were identified based on homologous detection and *de novo* searches. For homolog strategy, whole genome sequences were aligned with RepBase 21.01 (Jurka et al., 2005) using RepeatMasker (v4-0-6) program (www.repeatmasker.org). LTR_FINDER1.0.6 (Xu and Wang, 2007) was applied to identifying LTR retrotransposon elements to construct *de novo* repeat library, and genomic locations were also detected using RepeatMasker (v4-0-6).

Gene models were predicted based on the five closely related plant species of the rubber tree (*M. esculenta*, *R. communis*, *J. carcass*, *L. usitatissimum*and *P. trichocarpa*) and RNA-seq data. Amino acid sequences from these genomes were mapped using BLAT (Kent, 2002) with the rubber tree genome assembly to search for candidate protein-coding sequences, and then GeneWise (Birney et al., 2004) was performed to predict gene models. RNA-seq reads from bark, stem, seed, root, leaf and inflorescence tissues were mapped to the assembled *GT1* genome sequences with TopHat (Trapnell et al., 2009), and transcripts were determined by Cufflink (Trapnell et al., 2010); these evidences were subsequently integrated by GLEAN to generate conserved gene models. The integrated genes were compared with Cufflink and GeneWise results based on cultivar Reyan7-33-97 (Tang et al., 2016), and transcripts or functional genes were finally added to the GLEAN (Elsik et al., 2007) gene set. For function annotation, amino acid sequences were searched with known UniProt Consortium (2009) and, KEGG (Kanehisa and Goto, 2000) databases. InterProscan (Jones et al., 2014) were performed to annotate protein domains.

Analysis of Gene Family Evolution

Homologous genes from different species were combined using all *vs* all BLASTP. Gene families were clustered with OrthoMCL (Li et al., 2003) according to blast results. Gene family expansion and contraction were calculated using CAFÉ program based on gene cluster statistics. Homologous genes were detected by BLASTP (Altschul et al., 1990), and then synteny blocks were identified with MCscanX (Wang et al., 2012).

Population Genomic Analysis

The reads from 10 cultivated and 6 wild species were mapped to assemble the *GT1* genome using BWA (Li, 2013) with mem algorithm. The duplicated reads were removed using Picard tools (http://broadinstitute.github.io/picard/). SNP calling was performed using the Genome Analysis Toolkit (GATK) (McKenna et al., 2010). The phylogenetic tree of all individuals was reconstructed using Neighbor-Joining/UPGMA method, which was visualized using the software MEGA6 (Tamura et al., 2013). Population structure was speculated by ADMIXTURE (Alexander et al., 2009), which is based on a maximum likelihood method. We predefined the number of genetic clusters *K* from 2–7, and PCA analysis was performed using R language. The average pairwise diversity within a population ($\theta\pi$) and F_{ST} values were calculated on sliding windows of 100 kb. Considering that genomic regions under selection have a lowered diversity and reduced allele frequencies in the cultivated populations compared with the same region in wild ancestral populations, and thus $\log(p_{wild}/p_{cult}) > 1$ and $F_{ST} > 0.25$ were set as cutoff values to identify these regions.

RNA Isolation, Sequencing, and Assembly

Tissues from leaves, flowers, seeds, barks, roots and stems of *GT1* as well as the eleven other bark samples of the rubber tree, including 5 cultivated (W169, W589, W293, W53, W171) and 5 wild individuals (Y1187, Y379, Y809, Y4211, Y478) were collected in Jinghong City, Yunnan Province, China. The high-quality RNA was separately extracted and then sequenced on HiSeq 2500 platform. We filtered the low-quality reads by following the quality-control procedures used for the genome assembly. The transcriptome assembly for each rubber tree was generated using Trinity (version r20140717) (Grabherr et al., 2011) with default parameters. The gene expression levels were computed as the number of reads per kilobase of gene length per million mapped reads (FPKM) using RSEM software (Li and Dewey, 2011).

ACCESSION NUMBERS

All sequencing reads and genome assembly have been deposited in the NCBI and BIG Data Center under accession numbers PRJNA587314 and PRJCA001891, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by Yunnan Innovation Team Project and the start-up grant from South China Agricultural University (to L. G.).

AUTHOR CONTRIBUTIONS

L.-Z.G. and G.-H. L. conceived and designed the study; J. L., C. S., Ch. Sh., Y. W., C.-L. M., H.-B. Zh., G.-R. H.-T., X.-G. Zh. and S.-B. N. contributed to the sample preparation and genome sequencing; Ch.-Ch. Sh., G.-Y. F. and W. L. performed genome assembly; Y.T. performed flow cytometry experiments; Ch.-Ch. Sh., W.-B. Ch., G.-Y. F., Q.-J. Zh., and Y.Zh., W.L. performed genome annotation; Ch.-Ch. Sh., C. Sh., Q.-J. Zh., W. L., G.-Y. F., H.-F. L., Z.-Y. X., S.-T. Zh., Y. Y., X.-N. H., Sh.-J. H., W.-Q. L., M.-Q. L., J.-H. W., Y.-H. S., H. N., D. Zh., Y.-L. L.,Y. L.,K. L., J.-P. W. and Y.-N. J. performed data analyses; L.-Z.G., J. L., C. Sh., Ch.-Ch. Sh., Q.-J. Zh., and W. L. wrote the manuscript; L.-Z.G., X. L. X.-C. Zh., E. E. E. and X. L. revised the manuscript.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: October 16, 2019
Revised: October 16, 2019
Accepted: October 30, 2019
Published: December 12, 2019

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**:1655–1664.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Backhaus, R.A. (1985). Rubber formation in plants—a mini-review. Israel J. Bot. **34**:283–293.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Briere, C., Owens, G.L., Carrere, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature **546**:148–152.

van Beilen, J.B., and Poirier, Y. (2007). Establishment of new crops for the production of natural rubber. Trends Biotechnol. **25**:522–529.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. Genome Res. **14**:988–995.

Bowers, J.E. (1990). Natural rubber-producing plants for the United States. Proc. Natl. Acad. Sci. U S A **100**:1352–1357.

Bredeson, J.V., and Lyons, J.B. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat. Biotechnol. **34**:562–570.

Chan, H. (2000). Milestones in Rubber Research (Kuala Lumpur: Malaysian Rubber Board).

Chan, A.P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K.M., Redman, J., Chen, G., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. Nat. Biotechnol. **28**:951–956.

Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nat. Rev. Genet. **14**:645–660.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., and Eichler, E.E. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods **10**:563.

Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell **127**:1309–1321.

Dusotoit-Coucaud, A., Kongsawadworakul, P., Maurousset, L., Viboonjun, U., Brunel, N., Pujade-Renaud, V., Chrestin, H., and Sakr, S. (2010). Ethylene stimulation of latex yield depends on the expression of a sucrose transporter (HbSUT1B) in rubber tree (*Hevea brasiliensis*). Tree Physiol. **30**:1586–1598.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., and Weinstock, G.M. (2007). Creating a honey bee consensus gene set. Genome Biol. **8**:R13.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**:644–652.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463–467.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014).

InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **28**:27–30.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. **12**:656–664.

Lau, N.S., Makita, Y., Kawashima, M., Taylor, T.D., Kondo, S., Othman, A.S., Shu-Chien, A.C., and Matsui, M. (2016). The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. Sci. Rep. **6**:28594.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997 [q-bio.GN].

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**:323.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. **20**:265–272.

Lieberei, R. (2007). South American leaf blight of the rubber tree (Hevea spp.): new steps in plant domestication using physiological features and molecular markers. Ann. Bot. **100**:1125–1142.

Lin, T., Xu, X., Ruan, J., Liu, S., Wu, S., Shao, X., Wang, X., Gan, L., Qin, B., and Yang, Y. (2018). Genome analysis of *Taraxacum kok-saghyz* Rodin provides new insights into rubber biosynthesis. Natl. Sci. Rev. **5**:78–87.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**:1297–1303.

Ohno, S. (1970). Introduction in Evolution by Gene Duplication 1-2 (Berlin, Heidelberg: Springer).

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. **16**:1262–1269.

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., Theerawattanasuk, K., Nirapathpongporn, K., Romruensukharom, P., Tragoonrung, S., et al. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). Front. Plant Sci. **6**:367.

Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T., Sangsrakru, D., Theerawattanasuk, K., Rattanawong, R., Lekawipat, N., and Tangphatsornruang, S. (2017). *De novo* hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in *Hevea* species. Sci. Rep. **7**:41457.

Pritchard, J.K., Wen, X., and Falush, D. (2003). Documentation for STRUCTURE Software: Version 2.3 (Chicago: University of Chicago).

Priyadarshan, P.M., and Clement-Demange, A. (2004). Breeding *Hevea* rubber: formal and molecular genetics. Adv. Genet. **52**:51–115.

Priyadarshan, P.M., and Goncalves, P.D.S. (2003). *Hevea* gene pool for breeding. Genet. Resour. Crop Evol. **50**:101–114.

Rahman, A.Y., Usharraj, A.O., Misra, B.B., Thottathil, G.P., Jayasekaran, K., Feng, Y., Hou, S., Ong, S.Y., Ng, F.L., Lee, L.S., et al. (2013). Draft genome sequence of the rubber tree *Hevea brasiliensis*. BMC Genomics **14**:75.

Salse, J. (2016). Ancestors of modern plant crops. Curr. Opin. Plant Biol. **30**:134–142.

Sato, S., Hirakawa, H., Isobe, S., Fukai, E., Watanabe, A., Kato, M., Kawashima, K., Minami, C., Muraki, A., Nakazaki, N., et al. (2011). Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res. **18**:65–76.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. **16**:259.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. **30**:2725–2729.

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., An, Z., Zhou, B., Zhang, B., Tan, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants **2**:16073.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**:511–515.

UniProt Consortium. (2009). The Universal protein resource (UniProt) 2009. Nucleic Acids Res. **37**:D169–D174.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics **33**:2202–2204.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Webster, C., and Baulkwill, W. (1989). Rubber (Harlow, UK: Longman Scientific and Technical).

Wuyun, T.N., Wang, L., Liu, H., Wang, X., Zhang, L., Bennetzen, J.L., Li, T., Yang, L., Liu, P., Du, L., et al. (2018). The hardy rubber tree genome provides insights into the evolution of polyisoprene biosynthesis. Mol. Plant **11**:429–442.

Xia, E.H., Zhang, H.B., Sheng, J., Li, K., Zhang, Q.J., Kim, C., Zhang, Y., Liu, Y., Zhu, T., Li, W., et al. (2017). The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol. Plant **10**:866–877.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Yang, S.F., and Hoffman, N.E. (2003). Ethylene biosynthesis and its regulation in higher plants. Annu. Rev. Plant Physiol. **35**:155–189.

Zhang, D., Li, W., Xia, E.H., Zhang, Q.J., Liu, Y., Zhang, Y., Tong, Y., Zhao, Y., Niu, Y.C., Xu, J.H., et al. (2017). The medicinal herb *Panax notoginseng* genome provides insights into ginsenoside biosynthesis and genome evolution. Mol. Plant **10**:903–907.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal scale autopolyploid genomes based on Hi-C data. Nat. Plants **5**:833–845.